Procesamiento del lenguaje Natural, nº 41 (2008), pp. 299-300 recibido 7-05-2008; aceptado 16-06-2008

MOSTAS: Un Etiquetador Morfo-Semántico, Anonimizador y Corrector de Historiales Clínicos

MOSTAS: A Morpho-semantic Tagger, Anonymizer and Spellchecker for Clinical Reports

Ana Iglesias, Elena Castro, Rebeca Pérez, Leonardo Castaño y Paloma Martínez
Departamento de Informática
Universidad Carlos III de Madrid
Avda. Universidad, 30
28911- Leganés (Madrid)
aiglesia@inf.uc3m.es

José Manuel Gómez-Pérez, Sandra Kohler y Ricardo Melero
iSOCO S.A.
C/ Pedro de Valdivia, 10
28006 - Madrid
jmgomez@isoco.com

Resumen: El sistema MOSTAS pre-procesa historiales clínicos con el objetivo de facilitar el posterior tratamiento de los textos y recuperación de información de los mismos. El sistema añade información morfo-semántica a los historiales, busca el significado de las siglas, acrónimos y abreviaturas que existen en los mismos y detecta conceptos biomédicos, utilizando para ello recursos biomédicos especializados (bases de datos, tesauros, un servidor de terminologías multilingüe en OWL, etc.). Además, MOSTAS es capaz de anonimizar y corregir los historiales clínicos.
Palabras clave: Etiquetador morpho-semántico; Anonimizador de textos; Corrector Ortográfico; Conversión de Siglas, Abreviaturas y Acrónimos biomédicos; Historiales Clínicos; Historiales Médicos.

Abstract: The MOSTAS (MOrpho-Semantic Tagger, Anonymizer and SpellChecker for biomedical texts) system preprocesses Clinical Reports in order to facilitate rear information retrieval of these texts. MOSTAS system annotates clinical reports with morpho-semantic information, applies abbreviation and acronyms conversions and detects biomedical concepts using specialized biomedical resources (databases, thesaurus, a multilingual terminology server, etc.). Moreover, MOSTAS is able to anonymize and correct the clinical reports.
Keywords: Morpho-semantic tagger; Anonymizer; SpellChecker; Abbreviation and Acronym converser; Clinical Reports; Medical Reports.

1 Introducción y Motivación

En la actualidad existe un gran interés en el análisis de textos en el dominio de biomedicina con el objetivo de dar soporte a la búsqueda de documentación científica, ayuda a la toma de decisiones y seguridad de pacientes (Leroy y Chen, 2005).

Hasta el momento, la mayoría de los investigadores en este campo trabajaban con textos en inglés y terminología médica en inglés, pero aún queda mucho por hacer en los textos biomédicos en otros idiomas debido a la falta de estándares completos que aúnen terminologías (Lu et al., 2006). Además, la mayoría de los investigadores trabajan con documentos bien-formados como artículos, libros o resúmenes médicos similares a los que se pueden encontrar en bases de conocimiento como por ejemplo MedLine[1]. Pero aún existen pocos trabajos que estudien las notas escritas por personal de los hospitales, donde se pueden encontrar siglas, abreviaturas y acrónimos, palabras biomédicas especializadas y otros símbolos o palabras no controlados ni recogidos en los recursos biomédicos (Jang, Song y Myaeng, 2006).

2 MOSTAS: Etiquetador, Anonimizador y Corrector de Textos Clínicos

El sistema MOSTAS trata de pre-procesar de forma automática historiales clínicos de un

[1] http://www.nlm.nih.gov/medlineplus/

ISSN: 1135-5948

hospital con el objetivo de facilitar el posterior tratamiento de los datos y recuperación de información de los mismos. MOSTAS es un sistema creado para el proyecto ISSE[2], donde más de 210.700 notas clínicas de un hospital de Madrid han sido procesadas.

La arquitectura de MOSTAS se puede dividir en cuatro grandes bloques dependiendo del tratamiento del texto que se haga: Analizador Morfo-semántico, Buscador de Términos Médicos, Anonimizador y Corrector Ortográfico (ver Figura 1).

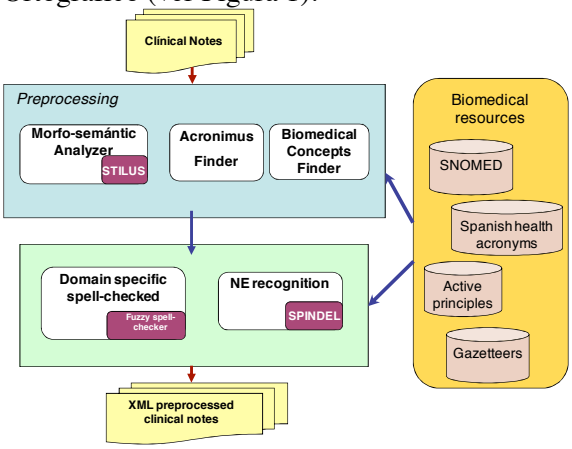

Figura 1: Arquitectura de MOSTAS.

El sistema MOSTAS recibe como entrada un conjunto de notas clínicas y proporciona como salida un documento XML con información morfo-semántica de las notas clínicas, buscando el significado de las abreviaturas y acrónimos en el texto, anonimizándolo y corrigiéndolo.

El analizador morfo-semántico de MOSTAS utiliza la herramienta STILUS[3], que detecta palabras en un diccionario general de español. Las palabras que no fueron reconocidas por STILUS se buscan en diccionarios de siglas, abreviaturas y acrónimos biomédicos (Nadeau, Turney y Stan, 2006). Si se encuentran, se almacenan en el documento XML los posibles significados que posee. En caso contrario, se busca su significado en diferentes recursos biomédicos mediante un servidor de terminologías[4] (ST) que posee información de metatesaurus como SNOMED[5], relacionados entre sí por un sistema de mapeo semántico. Para explotar la expresividad de las terminologías y facilitar el razonamiento, hemos dotado al servidor de un proceso de transformación de las distintas terminologías al lenguaje estándar W3C para la representación de conocimiento OWL. Posteriormente, teniendo en cuenta las palabras que aún no fueron detectadas en los diferentes recursos biomédicos, las notas clínicas se anonimizan utilizando SPINDEL (De Pablo et al., 2007), un buscador de entidades nombradas (personas, localizaciones y organizaciones). Por último, se parte de la hipótesis de que si aún existen palabras que no han sido reconocidas por los procesos anteriores, puede ocurrir que éstas estén mal escritas (algo que se ha observado que ocurre con frecuencia en este tipo de textos), por lo que se cuenta con un programa que busca similitud ortográfica de estas palabras mediante técnicas borrosas utilizando los recursos médicos especializados.

El documento XML con los textos clínicos etiquetados, corregidos y anonimizados por MOSTAS facilitará el trabajo posterior del tratamiento de los textos y recuperación de información de los mismos.

[2] ISSE: Interoperabilidad basada en Semántica para la Sanidad Electrónica. Proyecto PROFIT (FIT-350300-2007-75)

[3] http://stilus.daedalus.es/stilus.php

[4] ST desarrollado por iSOCO en proyecto ISSE

[5] http://www.snomed.org

Bibliografía

Jang, H., Song S.K., y Myaeng S.H. 2006. Semantic Tagging for Medical Knowledge Tracking Proc. 28th IEEE EMBS Annual International Conference.

Leroy, G. y Chen, H. 2005. Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. Journal of the American Society for Information Science and Technology, 56(1): 457-468.

Lu, W-H., Lin, R., Chan, Y-CH. y Chen, K-H. 2006. Overcoming Terminology Barrier Using Web Resources for Cross-Language Medical Information Retrieval. AMIA Annu Symp Proc., 519–523.

Nadeau, D., Turney, P.y Stan, M. 2006. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. 19th Canadian Conference on Artificial Intelligence. Québec City, Québec, Canada. June 7.

De Pablo, C., Martinez, J. L., Garcia-Ledesma, A., Samy, D., Martinez, P., Moreno-Sandoval, A., Al-Jumaily, H. MIRACLE Question Answering System for Spanish at CLEF 2007.